Explaining query modifications

An alternative interpretation of term addition and removal

Vera Hollink, Jiyin He*, and Arjen de Vries

Centrum Wiskunde en Informatica, Science Park 123, 1098 XG Amsterdam
{V.Hollink|J.He|Arjen.de.Vries}@cwi.nl

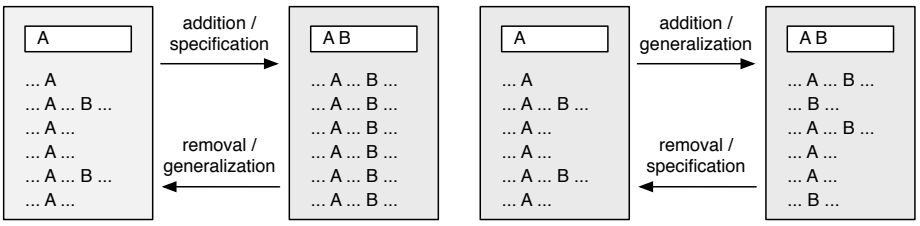
Abstract. In the course of a search session, searchers often modify their queries several times. In most previous work analyzing search logs, the addition of terms to a query is identified with query specification and the removal of terms with query generalization. By analyzing the result sets that motivated searchers to make modifications, we show that this interpretation is not always correct. In fact, our experiments indicate that in the majority of cases the modifications have the opposite functions. Terms are often removed to get rid of irrelevant results matching only part of the query and thus to make the result set more specific. Similarly, terms are often added to retrieve more diverse results. We propose an alternative interpretation of term additions and removals and show that it explains the deviant modification behavior that was observed.

1 Introduction

Searchers often need to modify their queries several times before their information needs are fulfilled or before they are confident that the collection does not contain any more relevant items. Research on query modifications studies transitions between consecutive user queries. Consecutive query pairs are classified based on the overlap in terms between the queries [3, 4, 7, 11, 12, 14, 15, 17, 21, 22]: term addition (e.g. from query `Maxima` to query `Maxima The Hague`), term removal (the opposite), and term substitution (e.g. from `Maxima The Hague` to `Maxima Dinner`).

Past research (e.g. [1, 3, 4, 7, 12]) identifies additions of terms with query specifications and removals of terms with generalizations. Although this interpretation of additions and removals is natural and generally accepted, to our knowledge there are no studies that verify whether specification and generalization is indeed what users intend when using these modifications. Correctly interpreting the intention of query modifications is not only a prerequisite to fully understand searching behavior, but also imperative when applying research on query modifications to improve search and search related tasks. For instance, in [22] the interpretation of query modifications is translated into four principles for reranking search results after various types of modifications. In [4] hyponyms are extracted from modifications. In [1, 14] models learned from modifications are used for query recommendation. These works implicitly make use of the

* supported by the Fish4Knowledge project funded by the 7th Framework Programme of the European Commission under grant agreement 257024.



a) Intersection-based interpretation b) Union-based interpretation

Fig. 1. Effect of term addition and removal on result lists. White boxes contain search queries; grey areas contain search results.

interpretation of term additions as specifications and removals as generalizations without verifying its validity.

The intuition behind viewing additions as specifications and removals as generalizations corresponds to the situation depicted in Figure 1a. Removing terms B from a query 'A B' keeps all documents that are about A and B and adds documents that are only about A, thus generalizing the result set. Conversely, adding terms B to query A selects the subset of the result set of A that is also about B: a specification of the result set. This interpretation is valid when all results that users view are relevant to all query terms. In other words, when the viewed results of a query 'A B' are the *intersection* of the viewed results of the individual query terms. Therefore, in the following we will refer to this interpretation as the *intersection-based interpretation*, the interpretation implicit in literature on query modifications.

In practice, result lists obviously do not always comply with the conditions of the intersection-based interpretation. The search engines that we consider in this work, like most state-of-the-art search engines, do not perform strictly boolean retrieval and thus do not only retrieve results containing all query terms. While such results tend to end up at the top of the result list (i.e. coordination level ranking like behavior [9]), below and sometimes between these are often results containing only some of the query terms. In this situation, term additions and removals can have different functions. As depicted in Figure 1b, when among the top ranked results of query 'A B' there are documents on only A and documents on only B (the *union* of the results of A and B), removing terms B results in a subset of the original result set containing only the documents on A. In this case, removing terms makes the top of the result set more specific instead of more general! Conversely, in case there are no (or only few) documents on both A and B, adding B to query A results in a mix of documents on A and documents on B: a generalization. This interpretation we will call the *union-based interpretation*.

The aim of this paper is to determine how well each of the two interpretations of term additions and removals can explain modification behavior of searchers. To this end, we examine cases of query modification extracted from query logs from three search contexts: web search, open domain image search (Flickr) and closed domain image search (a news photo agency). The first query in a pair of consecutive queries we call the *original query* and the second one *the modified query*. We investigate what motivated searchers to add and remove terms by studying the result sets of the original queries, under the assumption that users modify queries because of certain properties of the results retrieved by the original query that they like or dislike. To determine the

effects that the query modifications had on the results returned, we examine the result sets of the modified queries.

If users behave according to the intersection-based interpretation, additions will be used mainly when the result sets contain diverse results. That is, the user finds that the current result set contains too many items that are not relevant to his or her information need and adds terms to make the results more specific. Conversely, if removals are used as generalizations, they should be used predominantly when the result sets contain relatively coherent results.

The opposite effects would be predicted by the union-based interpretation. According to this interpretation, searchers add terms when the results of the original query are homogeneous, to get results on more diverse topics. In case the original result lists contain separate results for each of the query terms, and thus are not very coherent, searchers remove less important terms from a query to get rid of irrelevant results that are only about these terms.

We validate the extend to which searchers' modification behavior agrees with the predictions of the two interpretations by answering the following questions:

1. Do more coherent or less coherent result sets more often lead to term removals and term additions?
2. Do term additions and term removals increase or decrease the coherence of result sets?

As the result set of the original query can directly influence a user's modification behavior, the first question really investigates the motivations behind query modifications. The effects of modifications on the coherence (question 2) are under the control of the retrieval system rather than the users. However, these effects may indirectly still say something about the users' intentions, as users may have adapted their modification strategy to past experiences with the system, so that the modifications have the desired effects.

2 Related work

A large body of research has examined the usage frequency of query modifications. In most studies query modifications from query logs are classified on the basis of term overlap into addition/specification, removal/generalization, and substitution/reformulation [3, 4, 7, 10–12, 14, 17, 21, 22]. When queries are classified manually, sometimes not only term overlap but also the meaning of the queries is taken into account [14, 15, 19]. In [1] machine learning is used to learn the modification classes. Besides the three main modification classes, sometimes other types are examined as well, including lexical variations (e.g. changes from singular to plural forms) [3, 14, 19], lexical categories (part-of-speech tags) [2], and semantic relations [10]. For an overview we refer to [10]. The large majority of the studies of query modifications find that the most frequently used modification type is substitution/reformulation. Substitution/reformulation is used roughly twice as frequently as addition/specification, which is used roughly twice as frequently as removal/generalization [1, 3, 4, 10, 12, 15, 17, 19, 21].

Other aspects of query modifications have been researched, such as the time between queries and the relation between modifications and the occurrence of clicks [11].

To our knowledge, the present work is the first to relate query modifications to properties of result lists and to validate the standard interpretation of term additions and removals by investigating these properties.

In this work, we measure the specificity of queries through examination of the result lists. This type of specificity measures has been studied in other contexts as well and has shown to be effective in various applications. For instance, Cronen-Townsend and Croft [5] quantify the *ambiguity*, i.e., lack of specificity, of a query using the relative entropy between a query language model constructed from the top ranked retrieval results and the collection language model. Similarly, in [8] the ambiguity of a query is measured by comparing the tightness of the clustering structure of the documents associated with the query to a set of randomly drawn documents. Rudinac et al. [20] use the coherence of the top ranked results to predict if and how query expansion should be applied to a query. Even though for individual cases it may not always hold that more specific queries lead to more specific result lists, these studies have shown the effectiveness of result list-based specificity measures when averaged over many queries.

3 Method

To validate the intersection- and union-based interpretations of term removals and additions, we extract consecutive query pairs from a search log. The relation between these modifications and the coherence of the result sets is studied by examining for each pair the type of query modification and the coherence of the original and modified result sets. These steps are explained in more detail below.

The queries in the log are stemmed using a Porter stemmer. For each consecutive query pair, we classify the modification by determining whether, compared to the original query, in the modified query terms are added, removed, or substituted. In addition, we distinguish lexical variations, cases in which stemming makes the modified query identical to the original query (consecutive queries that were identical before stemming are conflated). Pairs without overlapping terms are classified as 'different'.

To understand when searchers use term additions and removals, we calculate the coherence of the result sets of the original queries (explained below). We compute the coherence of the result sets for all cases of term additions and all cases of term removals and compare these. In addition, to understand the effects of the modifications, we calculate the differences in coherence between the result sets of the modified and the original queries ($coherence(q_{modified}) - coherence(q_{original})$). We compare the differences of additions to the differences of removals.

Two measures are used to quantify the coherence of a result set. Let $D_q = \{d_i\}_{i=1}^{N}$ be a result set of documents retrieved with respect to query q. The first measure is the average pairwise similarity between the documents in D_q, defined as:

$$AvgSim(D_q) = \frac{\sum_{i<j \in \{1,\ldots,N\}} Sim(d_i, d_j)}{\frac{1}{2}N(N-1)}$$

where $Sim(d_i, d_j)$ is the cosine of the documents' term frequency vectors. A refinement of this measure is the coherence score proposed in [8], which has been shown to capture the topical coherence of a document set compared to a background collection:

$$Coherence(D_q) = \frac{\sum_{i<j \in \{1,\ldots,N\}} \sigma(d_i, d_j)}{\frac{1}{2}N(N-1)}$$

where

$$\sigma(d_i, d_j) = \begin{cases} 1 \text{ if } Sim(d_i, d_j) \geq \theta \\ 0 \text{ otherwise} \end{cases}$$

The threshold θ is determined by averaging the top $\tau\%$ similarity scores from document sets that are randomly sampled from a background collection.

To determine whether additions/removals can be interpreted as specifications/generalizations, the coherence score must provide a reliable estimation of specification and generalization. Therefore, in Section 5.1 we evaluate to what extend the coherence score agrees with the specificity of result lists as it is perceived by humans. In other words, we verify that an increase in coherence score is indeed perceived as a generalization and a decrease as a specification.

The union-based interpretation predicts that removals are mainly used when many results in the original result list do not contain all query terms and additions mainly when all results do contain all query terms. To validate these predictions, we examine the *coverage* of the search results of the original queries: the proportion of the results that contain all query terms. Let T_q be the set of query terms of query q and T_d the set of all terms in document d. We define the coverage of a set of search results D_q as:

$$Coverage(D_q) = \frac{|\{d|d \in D_q, T_q \subseteq T_d\}|}{|D_q|}$$

In the following, the coverage of a modification will refer to the coverage of the result list of its original query (and similarly for average similarity and coherence score).

4 Experimental setting

Data sets The first data set consists of the search logs of the commercial picture portal of a European news agency [10]. The portal provides access to more than 2M photos covering a broad domain. The log files record the search interactions of professional users (mainly journalists) accessing the picture portal. We use 10 months of search logs (October 2008 – July 2009), containing 1,094,620 queries in 520,507 sessions. Search sessions are identified using a log-in and a browser cookie and a time-out of 15 minutes.

The second data set consists of the queries submitted in the context of the iCLEF 2008 and 2009 evaluation campaigns [6]. Participants in the evaluation used the Flick-Ling system [18] to perform search tasks involving refinding specific images. Flick-Ling uses the Flickr API[1] to retrieve Flickr images. In total, the log files contain 56,894 queries from 7720 user sessions (participant/task pairs).

The third log file is from a major web search engine and consists of a set of 476,882 queries submitted in one day in 2006 (web_1 _day). To keep the number of requests issued to the search engine small (see Section 'Retrieval systems'), we randomly sample

Table 1. Number of query pairs in the data sets.

	News	iCLEF	web_20k	web_1_day
all	556,007	49,174	20,000	199,972
2 terms	282,039	15,713	4,842	not used
$\geq$ 2 terms	355,660	44,132	17,659	not used

from this data set 10,000 cases of term additions and 10,000 cases of term removals (web_20k). The number of query pairs in each data set can be found in Table 1.

Searchers may apply term additions and removals to queries of different lengths. In particular, term removal can only be used when the original query has at least 2 terms. As query length potentially effects the coherence of the result sets, this difference could interfere with our results. To compensate for this, we look at two subsets of the set of query pairs: 1) query pairs where the original query has *exactly* 2 terms (all effects of query length eliminated) and 2) query pairs where the original query has *at least* 2 terms (all cases where both additions and removals are possible). For removals, the last subset equals the entire set of removals.

Table 2 shows that the relative frequencies of the modifications in our data sets are in line with the proportions found in other studies (see Section 2). This is an indication that, with respect to query modifications, the examined data sets are comparable to the ones used in other studies.

Table 2. Relative frequency of the various query modifications.

Modification	News	iCLEF	web_1_day
substitution	0.12	0.36	0.25
addition	0.08	0.27	0.14
removal	0.04	0.18	0.06
lexical variation	0.01	0.02	0.03
different	0.75	0.17	0.53

Retrieval systems The search logs record queries and clicked documents, but not the returned results. To get the result lists we rerun the queries. The original search engine of the News photo agency does not provide an API. Therefore, the queries are rerun on a locally stored subset of the image collection, consisting of 2,247,035 images from the same time period as the log files. We index the metadata of the images in the subset using the Lemur toolkit[2]. We retrieve images using the query likelihood model with default parameter settings. For the second data set, we submit the queries to the Flickr API exactly as they were submitted by FlickLing (in some cases FlickLing translates the query before submitting it [18]). The retrieved images are associated with the same information that was visible to the iCLEF participants: title and tags. For the third data set, we submit the queries to the Bing API[3] to retrieve Web pages. As in the search interface from which the logs originated, each retrieved page is associated with its url, title, and a snippet of the text surrounding the query in the page.

[1] http://www.flickr.com/services/api/

[2] http://www.lemurproject.org/

[3] http://www.bing.com/developers

We always retrieve the top 16 documents as our result set to represent the set of results that the searcher viewed before deciding to make a modification. This number was chosen because the search engines in this study typically show between 8 and 20 search results per result page and users tend to browse one or two result pages [13].

Due to differences in retrieval systems and collections, the result sets we obtain with respect to a query may differ from what the searcher saw when she/he issued the query. Further, searchers may have examined more or less than 16 results. Despite these anticipated differences, we expect that in most cases the result sets have the same properties, e.g. queries that gave many results will still give many results and ambiguous queries will still give diverse results.

Parameter settings Following [8], we set the parameter τ in the coherence score to 0.05%. As a result, the threshold θ obtained from the background collections becomes 0.05, 0.44, and 0.18 for the News, iCLEF, and web_20k data set respectively. Since we do not have the document collections for the iCLEF and web data sets, we pool over all documents retrieved in response to the queries as the background collections. Despite the fact that these documents are not random samples of the collections (they are associated with the queries), given the large number of queries used and the relatively small number of documents retrieved per query, we expect that the pooled collections will not be too biased towards certain queries.

5 Results

5.1 Reliability of the coherence score

Before continuing, we assess the reliability of the coherence score. We determine whether the specificity of result sets as predicted by the measure is consistent with the specificity as perceived by humans.

We sampled 120 modifications from the data sets. For each modification we showed the result set of the original and the modified query (without the queries) to human judges. Not all top 16 results were shown but 8 results that were sampled randomly from the top 16. This reduced the time the judges needed per modification and thus enables us to get judgements for more modifications. The judges were asked to indicate whether the first result set was *more specific*, *equally specific*, or *less specific*, than the second result set, or that the result sets were *incomparable*.

If the measure is reliable, the label *more specific* for the original result set should be associated with a strong negative coherence difference between the modified and the original query, *less specific* with a strong positive coherence difference and *equally specific* with a small coherence difference (around 0). Therefore, we sampled the modifications in such a way that 40 modifications had a strong negative coherence difference (randomly sampled from the modifications with the lowest 5% coherence score differences, e.g. for the web data set additions with differences between -1 and -0.52), 40 modifications a strong positive coherence difference (from the highest 5%, for web additions between 0.43 and 1) and 40 modifications a small coherence difference (from the middle 25%, for web additions between -0.04 and 0.04). Within these groups we ensured equal numbers of removals and additions and equal numbers from the News

Table 3. Confusion matrix of the majority labels assigned by human judges and the labels assigned on the basis of coherence score. Matching assignments are shown in bold.

	Coherence score		
Majority	strongly negative	around zero	strongly positive
More specific	**23**	3	0
Equally specific	3	**26**	6
Less specific	0	5	**20**

and the web data sets. The iCLEF data set was not used to reduce the workload of the judges.

The samples were divided randomly in two sets. Six judges each labeled one set (60 pairs of result sets), so that each pair of result sets was labeled by 3 judges. The order of the pairs as well as the order of the result lists within each pair were randomized. The ground truth label for each pair was determined by majority voting. Samples for which no majority was found (10 pairs) or for which the majority label was 'incomparable' (24 pairs), were left out of the analysis.

For each of the remaining 86 samples, we compare the ground truth label to the class assigned on the basis of the coherence score (strongly negative, strongly positive, around zero). Results are shown in the confusion matrix in Table 3. The majority label and the coherence score label agreed on 80% of the samples. Cohen's kappa was found to be κ=0.70, a substantial agreement [16]. From this we conclude that the coherence score provides a fair representation of specificity as it is perceived by humans.

5.2 Validation of the two interpretations

We now continue to validate the predictions of the intersection-and union-based interpretations, answering the two questions posed in the introduction. We compare the coherence of the additions and removals found in the data sets described in Section 4.

Do more coherent or less coherent result sets more often lead to term removals and term additions?

Table 4 shows the average coherence and average pairwise similarity of the results of the original query of additions and removals. The results are in line with the expectations of the union-based interpretation and exactly the opposite of what would be expected based on the intersection-based interpretation: users add terms when the search results are already relatively coherent and remove terms when the results are relatively incoherent. This finding is independent of the query length and the way we measure coherence (average similarity or coherence score). The differences between the scores of additions and removals are all significant at $p<0.01$ (Wilcoxon rank sum test).

The average coverage of additions and removals is shown in the right part of Table 4. According to the union-based interpretation, the low average coherence associated with removals, is caused by a relatively large number of search results that do not contain all query terms (low coverage). The fact that additions often occur when coherence is high, is attributed to a large number of search results that do contain all query terms. The table confirms these predictions on all three data sets: additions are associated with a much higher average coverage than removals. In fact, as shown in Figure 2, terms are

Table 4. Coherence score, average pairwise similarity, and coverage of the result set of the original query averaged over all cases of term addition and over all cases of term removal. $\gg/\ll$ indicates significantly larger/smaller with p-value <0.01 using the Wilcoxon rank sum test.

Data set		Coherence			Average similarity			Coverage		
		add		remove	add		remove	add		remove
News	all	0.65	$\gg$	0.56	0.56	$\gg$	0.52	0.90	$\gg$	0.29
	2 terms	0.66	$\gg$	0.57	0.56	$\gg$	0.52	0.78	$\gg$	0.40
	$\geq$ 2 terms	0.66	$\gg$	0.56	0.56	$\gg$	0.52	0.73	$\gg$	0.29
iCLEF	all queries	0.94	$\gg$	0.71	0.32	$\gg$	0.29	0.80	$\gg$	0.39
	2 terms	0.94	$\gg$	0.73	0.34	$\gg$	0.27	0.81	$\gg$	0.51
	$\geq$ 2 terms	0.94	$\gg$	0.71	0.35	$\gg$	0.29	0.75	$\gg$	0.39
web_20k	all queries	0.68	$\gg$	0.64	0.28	$\gg$	0.27	0.69	$\gg$	0.35
	2 terms	0.70	$\gg$	0.58	0.29	$\gg$	0.25	0.80	$\gg$	0.61
	$\geq$ 2 terms	0.73	$\gg$	0.64	0.30	$\gg$	0.27	0.64	$\gg$	0.35

Table 5. Difference in coherence score, average pairwise similarity, and coverage between the result set of the modified query and the result set of the original query averaged over all cases of term addition and over all cases of term removal. $\gg/\ll$ indicates significantly larger/smaller with p-value <0.01 using the Wilcoxon rank sum test.

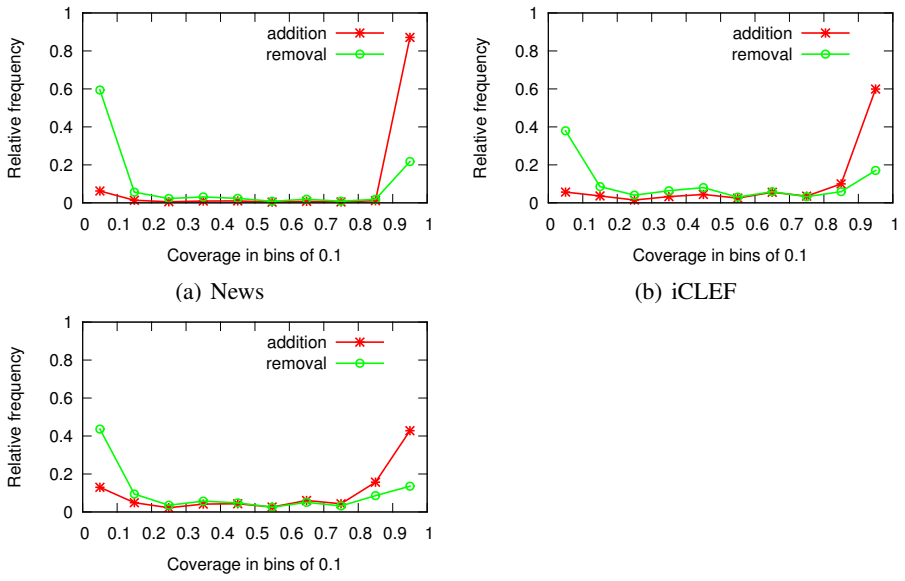
Data set		Coherence diff.			Average similarity diff.			Coverage diff.		
		add		remove	add		remove	add		remove
News	all	-0.035	$\ll$	0.072	-0.016	$\ll$	0.034	-0.449	$\ll$	0.554
	2 terms	-0.031	$\ll$	0.078	-0.012	$\ll$	0.034	-0.455	$\ll$	0.601
	$\geq$ 2 terms	-0.031	$\ll$	0.072	-0.013	$\ll$	0.034	-0.424	$\ll$	0.554
iCLEF	all queries	-0.138	$\ll$	0.186	-0.012	$\ll$	0.025	-0.282	$\ll$	0.323
	2 terms	-0.151	$\ll$	0.190	-0.029	$\ll$	-0.015	-0.296	$\ll$	0.406
	$\geq$ 2 terms	-0.148	$\ll$	0.186	-0.033	$\ll$	0.025	-0.278	$\ll$	0.323
web_20k	all queries	-0.013	$\ll$	0.039	0.002	$\ll$	0.010	-0.320	$\ll$	0.337
	2 terms	-0.024	$\gg$	-0.080	-0.000	$\gg$	-0.042	-0.384	$\ll$	0.256
	$\geq$ 2 terms	-0.054	$\ll$	0.039	-0.014	$\ll$	0.010	-0.321	$\ll$	0.338

added predominantly when all results contain all query terms (coverage is 1). Terms are mainly removed when there are no results containing all query terms (coverage is 0).

Do term additions and term removals increase or decrease the coherence of result sets?

The results of the modifications on the coherence of the result sets are shown in Table 5. Again, our results are in agreement with the union-based interpretation rather than the intersection-based interpretation: in 8 out of 9 (sub)data sets, additions are associated with significantly smaller average differences than removals. Additions tend to decrease coherence, while removals tend to increase coherence. For completeness, Table 5 also shows the effects of the modifications on the coverage of the result sets. Naturally, adding terms decreases coverage and removing terms increases coverage.

We conclude that the behavior observed in the usage logs is not in line with the interpretation of query modifications that is common in the query log studies discussed



(a) News (b) iCLEF

(c) web_20k data

Fig. 2. Distribution of the coverage of the original query before term additions and term removals.

before. We will now take a closer look at some modifications to understand better what causes this discrepancy.

An example from the News logs of a removal that made the results more specific is the following. A searcher first used the query `cleaner of the year`. The top 16 results contained 8 results with as only query term `year` (coverage 0.5). The user then modified the query into `cleaner`, apparently to remove the irrelevant `year` results. This behavior is consistent with the union-based interpretation: searchers remove terms B from a query 'A B' to get rid of irrelevant results that are only about B. A similar removal example taken from the web_20k logs transforms `shoulder pain erand numbness in index finger` to `numbness in index finger`. When the first query did not retrieve any results containing all query terms (coverage 0), the user presumably decided to remove the shoulder pain results and focus on the numbness. An example from the web_20k logs of an addition that made the query less specific is from `bible` to `bible tora coran`. Again, this example is consistent with the union-based interpretation.

As discussed, Figure 2 shows that in the majority of cases searchers' modification behavior is consistent with the union-based interpretation. However, there is also a non-negligible number of cases where the intersection-based interpretation better explains the observed user behavior: 6% (iCLEF) to 13% (web_20k) of the additions have a coverage of 0.1 or less and 14% (web_20k) to 22% (News) of the removals have a coverage above 0.9. Examples of removals where the intersection-based interpretation is likely to apply are from `jack daniel whiskey` to `whiskeys` and from `alltell wireless` to `alltell`. These examples have high original coverage and

decreasing coherence and appear to be intended as generalizations. Additions consistent with the intersection-based interpretation are, for instance, `dallas` to `dallas visitors bureau` and `grey flannel` to `grey flannel lotion`. This shows that both interpretations are needed to effectively explain searchers' modification behavior. For applications that make use of modifications, this suggests that for optimal performance both interpretations should be taken into account.

6 Conclusion

The main contributions of this paper are twofold: 1) we presented a method to study the relation between query modifications and the coherence of result sets, and 2) we determined to what extend the widely used intersection-based interpretation of term additions as specifications and removals as generalizations is valid.

Our experiments show that additions are often not used to specialize result sets and removals often not used to generalize result sets. In fact, judging from the observed result list coherences, in the majority of the cases the modifications appear to have the opposite functions: terms are removed to get rid of irrelevant results matching only part of the query and terms are added to retrieve more relevant results: the union-based interpretation.

Although in hindsight the union-based interpretation may look as natural as the intersection-based interpretation, it is currently not taken into account in applications that make use of modifications in log analysis or for improving search. By identifying additions with specifications and removals with generalizations, such applications implicitly assume that the intersection-based interpretation is always valid. Our findings imply that for log analysis this simplification may lead to a biased view on the intentions behind query modifications. Applications that use modification to optimize search strategies can potentially also be improved by distinguishing cases where the intersection-based and the union-based interpretation apply. For example, the reranking principles defined in [22] may be refined by making separate principles for each of the interpretations.

The provided measures (average similarity, coherence, and coverage) are promising measures to distinguish between the two interpretations: when their values are low the union-based interpretation of term removal is likely to apply while high values point at the intersection-based interpretation and vice versa for additions. The next step will be to determine which (combination of) measures can most accurately predict the real intentions of searchers. By asking searchers in an interactive experiment for their motivations when making modifications, we will validate whether the proposed explanations agree with actual motivations of users and measure the predictive power of the measures. Once we can accurately distinguish a user's motivation for making a modification, the distinction may be applied to tailor search.

Bibliography

[1] Boldi, P., Bonchi, F., Castillo, C., Vigna, S.: Query reformulation mining: models, patterns, and applications. Information Retrieval 14(3), 257–289 (2010)

[2] Bozzon, A., Chirita, P.A., Firan, C.S., Nejdl, W.: Lexical analysis for modeling web query reformulation. In: SIGIR'07. pp. 739–740 (2007)
[3] Bruza, P., Dennis, S.: Query reformulation on the internet: empirical data and the hyperindex search engine. In: RIAO'97. pp. 488–499 (1997)
[4] Costa, R.P., Seco, N.: Hyponymy extraction and web search behavior analysis based on query reformulation. In: IBERAMIA'08 (2008)
[5] Cronen-Townsend, S., Croft, W.B.: Quantifying query ambiguity. In: HLT '02. pp. 104–109 (2002)
[6] Gonzalo, J., Peinado, V., Clough, P., Karlgren, J.: Overview of iCLEF 2009: exploring search behaviour in a multilingual folksonomy environment. In: CLEF'09, pp. 13–20 (2010)
[7] He, D., Göker, A., Harper, D.J.: Combining evidence for automatic web session identification. Information Processing and Management 38(5), 727–742 (2002)
[8] He, J., Larson, M., De Rijke, M.: Using coherence-based measures to predict query difficulty. In: ECIR'08. pp. 689–694 (2008)
[9] Hiemstra, D.: Term-specific smoothing for the language modeling approach to information retrieval: the importance of a query term. In: SIGIR'02. pp. 35–41 (2002)
[10] Hollink, V., Tsikrika, T., De Vries, A.P.: Semantic search log analysis: a method and a study on professional image search. JASIST 62(4), 691–713 (2011)
[11] Huang, J., Efthimiadis, E.N.: Analyzing and evaluating query reformulation strategies in web search logs. In: CIKM'09. pp. 77–86 (2009)
[12] Jansen, B.J., Booth, D.L., Spink, A.: Patterns of query reformulation during web searching. JASIST 60(7), 1358–1371 (2009)
[13] Jansen, B.J., Spink, A., Pedersen, J.O.: An analysis of multimedia searching on AltaVista. In: MIR'03. pp. 186–192 (2003)
[14] Jones, R., Fain, D.C.: Query word deletion prediction. In: SIGIR'03. pp. 435–436 (2003)
[15] Jörgensen, C., Jörgensen, P.: Image querying by image professionals. JASIST 56(12), 1346–1359 (2005)
[16] Landis, J., Koch, G.: The measurement of observer agreement for categorical data. Biometrics 33(1), 159–174 (1977)
[17] Özmutlu, H.C.: Markovian analysis for automatic new topic identification in search engine transaction logs. Applied Stochastic Models in Business and Industry 25(6), 737–768 (2009)
[18] Peinado, V., Artiles, J., Gonzalo, J., Barker, E., Ostenero, F.L.: FlickLing: a multilingual search interface for Flickr. In: CLEF'08 (2008)
[19] Rieh, S.Y., Xie, H.: Analysis of multiple query reformulations on the web: the interactive information retrieval context. Information Processing and Management 42(3), 751–768 (2006)
[20] Rudinac, S., Larson, M., Hanjalic, A.: Exploiting result consistency to select query expansions for spoken content retrieval. In: Advances in Information Retrieval, LNCS, vol. 5993, pp. 645–648 (2010)
[21] Whittle, M., Eaglestone, B., Ford, N., Gillet, V.J., Madden, A.: Data mining of search engine logs. JASIST 58(14), 2382–2400 (2007)
[22] Xiang, B., Jiang, D., Pei, J., Sun, X., Chen, E., Li, H.: Context-aware ranking in web search. In: SIGIR'10. pp. 451–458 (2010)